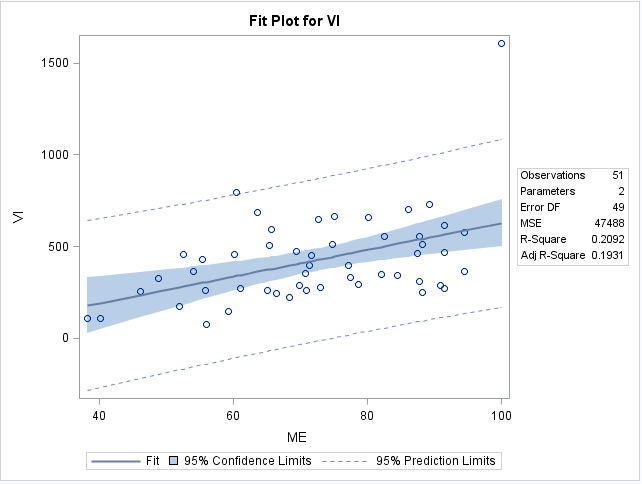
Remy Lagrois

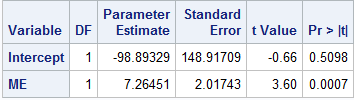
HW 11

Section 402

1.  a)

The ME and VI do appear to be linearly correlated. The slope is positive meaning that as the percentage of people who live in a metropolitan area increases so does the rate of violent crime. However there seem to be other major factors that influence the crime rate as the R2 value is 0.2092 meaning that the metro population only explains 20.92% of the observed crime rate.

b) The regression equation is y = -98.893 + 7.265x where ‘y’ is the violent crime rate and ‘x’ is the percent of people living in a metro area.

c) Yes the slope is significantly different than zero and is positive.

The slope was found to be 7.26451 with a p value of 0.0007 meaning we reject the null hypothesis that the p-value is zero.

d) The 95% confidence interval for the slope is 3.210 to 11.319. This further reinforces that the slope is both not zero and positive since the interval does not contain zero nor any negative numbers.

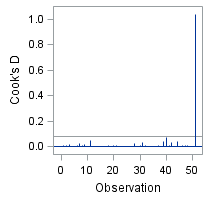
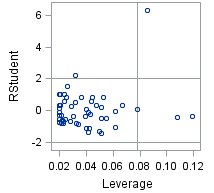
2. a) The four assumptions of linear regression are that each measurement is independent, the residuals have equal variance, the residuals are normally distributed, and that there is a linear relationship between the dependent and independent variables.

b) If the assumptions aren’t violated then the plot should be a random cloud with no discernable shape. There should also be a roughly equal number of points above and below the line y = 0.

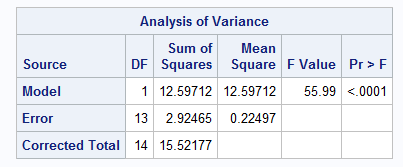
c) The QQ plot should be a straight positively sloped line going right up the middle of the plot.

d) The spread of the points would be uneven and may form a cone/fan type shape instead of a random cloud.

e) Yes there are a few high leverage values in this data set.

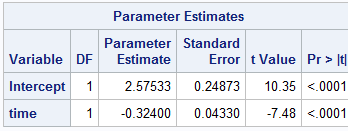


The plots above show there are 3 points with quite high leverage. The one that sticks out the most is Washington DC represented by the point above 0.08 leverage and greater than 6. This shows that in addition to its high leverage it is also very far from its predicted value. It also has high influence. The other two points with very high leverage are much closer to their predicted values and a probably the states of Maine and West Virginia which have low proportions of metro population but whose crime rate is very close to the predicted one.

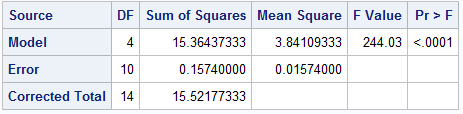


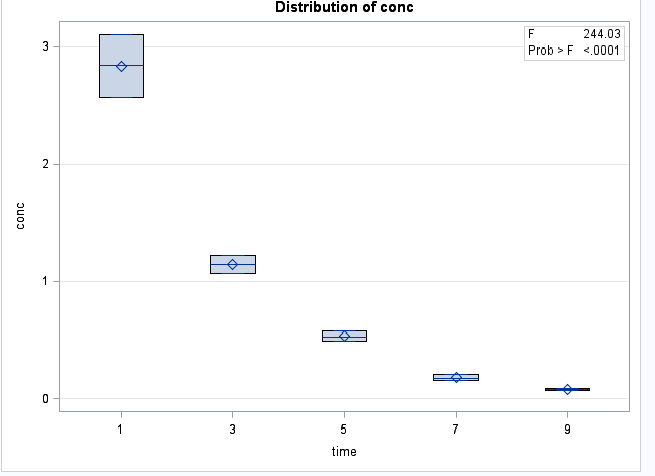
3. a)

The ANOVA table shows there is significant difference between the different times, given the p-value of 0.0001, so using some type of regression is appropriate.



The estimated slope is -0.324 and the p-value shows that the slope is significantly different from zero. As time passes the chemicals decay and the concentration decreases. Time can also be seen to be a significant predictor of concentration as the R2 is rather high at 0.811.

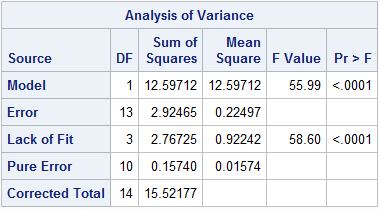
b)

The low p-value of <0.0001 allows us to reject the null hypothesis that all of the final concentrations are equal. Instead we accept that at least one is different from the other. Looking at the box plot provides confirmation of this and suggests that as time goes on the concentration decreases. However it would also appear that it may not be a linear relationship and the rate of decay slows over time.

c) The sum of squares total uses the difference between the mean of all concentrations and each observed concentration squared and all added together. So since the data is the same in both the ANOVA and regression model the sum of squares total will also come out to be the same.

d) In ANOVA the SS(error) is a measure of variation within each group (as in variation within time 1, time 3, etc). Regression has the within group variation but also includes the difference between the prediction and observed so the SS(error) is greater.

e) Regression and ANOVA have different goals overall. ANOVA is simply looking at whether or not there is a difference between groups. Regression does this as well but also tries to make a prediction by determining an equation for a line that most closely matches the data. The accuracy of that prediction must be tested in order to determine if it has any value.

f)

The lack of fit has a very low p-value of <0.0001. This means that we reject the null hypothesis that our model is a good fit and instead we must accept that our model needs more terms or variables in order to get a good fit. This is consistent with what we saw with the box plot and residual plots from earlier. While time is a good predictor in that concentration gets lower over time it doesn’t allow us to make accurate predictions for what the exact concentration will be using linear regression.